e-Library Science Research Journal
Vol.1,Issue.9/July. 2013 ISSN : 2319-8435

Research Papers

DATA MINING AS AN EMERGING TECHNIQUE: IT MAY EXPAND THE CREATIVITY OF LIBRARIES

P. S. RAJPUT AND MANOJ KUMAR CHOUDHARY

Asstt. Librarian University College of Science, Mohan Lal Sukhadia University, Udaipur, (Raj.)
Defence Institute of Advance Technology, Girinagar, Pune.

Abstract

Data Mining (DM), the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help the Libraries and Information Centers to focus on the most important information in their data warehouses. Data mining popularly known as Knowledge Discovery in databases is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases. The present study attempts to know the meaning of data mining, importance and its process. Explains the different steps of data mining and its implementation in libraries. Finally suggests the guideline to choose a data mining system.

KEYWORDS

Data mining; Knowledge discovery; Data warehouse, Data mining Techniques.

1.INTRODUCTION

In this age of Information Technology our capabilities of generating and collecting data have been increasing rapidly every day. Data Mining is a promising new technology in database systems and applications which assist us in transforming the vast amounts of data into useful information and knowledge. The term data mining that emerged during 1980s, has made great strides during the 1990s and is expected to continue to flourish into the new millennium. It is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories.

University libraries have the good collection of information and in e-library there is organized collection of information, which serves a rich resource for its user communities. Information resources appear in the library collection, users locate required information using catalogue search system and bibliographic databases. The vast data stored in the databases of traditional and digital libraries represent the behavioral patterns of two important constituents i.e. library staff and library users.

In the case of library staff mining available acquisitions and bibliographic data could provide important clue to understanding and enhancing the effectiveness of the library's own internal functions. Mining user data for knowledge about what information library users are seeking, whether they locate what they hunt for, and whether their queries are satisfied these kinds of information can have strategic utility within the large organization in which the library is situated.

2.WHAT IS DATA MINING

Data Mining refers to extracting or mining knowledge from large amounts of data. Just like gold mining from rocks and sand, data mining should have been more appropriately named as 'knowledge mining from data.' Since the term reflects emphasis on mining from large amounts of data, it is termed as "Data Mining". Data mining is a term used mainly in computer science. Sometimes it is also called knowledge discovery in databases (KDD). Data mining is about finding similarities in large sets of data. It is about discovering patterns in large sets of data. Very often, such data is stored in some form of database. Some commonly used algorithms can be classified as pattern-recognition, or Neural Network. According to Wikipedia Data mining is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management

In short data mining means extraction of interesting information or patterns from data in large databases. Data mining is an essential step in the process of knowledge discovery in databases. The alternative names of data mining are Knowledge Discovery in Databases (KDD), Knowledge extraction, Data pattern analysis, Data archaeology, Data dredging, Information harvesting etc.

3.IMPORTANCE OF DATA MINING

"Necessity is the mother of invention". The major reason that data mining has attracted a great deal of attention in the information industry in recent years the wide availability of huge amounts of data and the imminent need for running such data into useful applications ranging from business management, production control and market analysis to engineering design and science exploration.

Data mining can be viewed as a result of the natural evolution of information technology. The steady and amazing progress of computer hardware technology in the past three decades has led to large supplies of powerful computers, data collection equipment and storage media. The data warehouse, a database for storing data includes data cleaning, data integration and On-line Analytical Processing (OLAP). The fast growing tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful tools. Data explosion problem evolved from automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories. We are drowning in huge data, but starving for knowledge. Data mining has presented a solution to this situation.

4.PROCESS OF DATA MINING

Acquiring knowledge from data warehouse is a concrete process. The major steps are mentioned below.

1. Data gathering, e.g., data warehousing
2. Data cleansing: eliminate errors and/or inconsistent data
3. Data selection: where data relevant to the analysis task are retrieved from the database.
4. Data transformation
5. Data mining: an essential process where intelligent methods are applied in order to extract data
6. Pattern evaluation: to identify the truly interesting patterns representing knowledge based on some interesting measures
7. Knowledge presentation: visualization techniques used to present the mined knowledge to the user

5.DATA MINING AND LIBRARIES

This part of the study explores the ways data mining can be useful in the field of Library and Information Science. As per the fifth law of Library Science "Library is a Growing Organization," so the volume of the library data is also growing with an enormous rate. The library creates data sources appropriate for bibliomining before it obtains new information. Sources (e.g. books, reference tools, databases, electronic access etc.) a librarian assesses the needs of the existing collection in light of available and upcoming publications.

Bibliomining is derived from the term "bibliometrics" and "data mining", as the goal is to take advantage of the social networks that power bibliometrics and user based data mining through a single data warehouse. It is defined as "the combination of data mining, bibliometrics, statistics and reporting tools used to extract patterns of behavior-based artifacts from library system". It is the application of statistical and pattern-recognition tools to large amounts of data associated with library systems in order to aid decision-making or justify services. The bibliomining process consists of:

Determining areas of focus;
Identifying internal and external data sources;
Collecting, and cleaning the data into a data warehouse;
Selecting appropriate analysis tools;
Discovery of patterns through data mining and creation of reports with traditional analytical tools; and
Analyzing and implementing the results.

In order to locate works in the library, users rely on the OPAC. Libraries often examine users comments and surveys to assess users satisfaction with these tools. Therefore, librarians may wish to examine the artifacts of those searches for problem areas instead of replacing on users comments and surveys in order to improve the user experience. When upgrading or changing library system interfaces, librarians can explore these patterns of common mistakes in order to make informed decisions about system improvement.

5.1.Data Mining Techniques Improve Library System

Assume that a library is determined to fulfill the following objectives:

Increase books borrowing rate
Attract more number of users to borrow books
Assist library professionals in making policy on the acquisition of duplicate copies and new publications.

Then library may perform activities such as fellow:

5.1.1 Acquisition: As per third law of Library Science "Every Books Its Reader" by applying data mining in the library data it can be easily find out the required contents that are necessary to acquired next. This will reduce the work of library staff related to the acquisition as well as the efficient use of budget allocated to the library.

5.1.2 Classification: It replaces the manual classification of content of library with the computer assisted classification, so that classification task can be accomplished even by less skilled persons timely and efficiently.

5.1.3 Reference Service: Searching of information as the data of library is continuously growing with an exponential rate and the main problem is how one can use the required information from the large amount of redundant information of the library. This can be possible by applying data mining techniques. So one can say that the data mining is the future of reference service.

5.1.4 Sequence Analysis: Sequence analysis uses statistical analysis to identify unlinked documents that users are likely to read together. It also examines the paths that users follow when searching for information.

5.1.5 Summarization: Though machine generated abstracts are inferior to human-generated ones in terms of readability and content yet they can be useful for helping users decide what items they need. Abstract-generating software typically works by identifying significant words or phrases based on position within documents association with critical phrases

5.2 Library Management through Data Mining

The implementation and use of data mining system for libraries and information centers management can be better understood from the example and the figure given below. Consider a library where books database includes more than 150000 volumes of books and the borrowing history of 10000 users who have borrowed books from the library. From this database we have developed a report system to represent the relationship between books and students. Then we use data mining association rule to analyze the books of the same cluster borrowed by the students. A report chart of books will be produced, so that the library professionals will find references to the books for making a decision. Additionally, the library collection may be used more effectively.

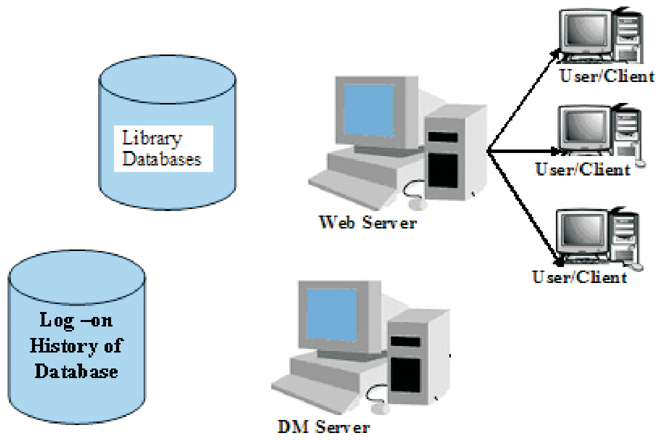

Figure 1: Design of MD System for Libraries

6.GUIDELINES FOR COMPETENT DATA MINING SYSTEM

The following may be the best guidelines for data mining system.
Data types: text, multimedia etc.
System issues: UNIX, Microsoft windows Linux.
Data sources: data formats like, ODBS, for OLAP, various data supporting standard
DM function and methodologies: Multiple function and methods per functions provide the user with greater flexibility and analysis power.
Scalability: Scalability of query (1) Row (2) column
Visualization: A picture is worth a thousand words. So visualization tools may strongly influence the data mining system
DM query language and graphical interface: SQL, OLEDB (Object linking and embedding database)

7.CONCLUSION

Since the amount of information and knowledge published both electronically and physically are increasing at a tremendous speed, to retrieve a particular data from the huge amount of data is very difficult. Data mining is boon to the information professionals to extract the potential data from the vast amount of information. There are lot of private companies and institutions of data mining available on-line and we must choose suitable data mining software and implement the same successfully to satisfy the data needs of the users.

REFERENCES

1. Hand, D., Mannila, H., and Smyth, P. (2001). Principles of Data Mining, MIT Press, Cambridge, pp.4-35.
2. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
3.Dhiman, A.K. (2003), "Data Mining and its used in Libraries", Proceeding of CALIBER, Ahmedabad, February 13-15, 2003, pp. 568-74.
4.http://en.wikipedia.org/wiki/Data_mining
5.Kantardzic, M. (2003). Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons, Cambridge, pp. 15-60.
6.Mukhopadhyay, B. and Mukhopadhyay, S. "Application of Data Mining Techniques for Library Management Information System", available at: http://ir.inflibnet.ac.in/dxml /bitstream/handle/1944/226/cali_57.pdf?sequence=1 (accessed 23 April 2012).
7.http://www.thearling.com/text/dmwhite/dmwhite.htm